Element of Data Processing Assignment 2

Group W{04}G{8}

Tania Candra – 1364261
Alex Lai – 1406509
Michael Xianhe - 1461175

1. Executive Summary

Our report investigates potential factors that may influence book ratings at a U.S. online bookstore, analyzing data on user profiles, book details, and ratings. After pre-processing and cleaning the data where necessary, we applied machine learning techniques such as Decision Trees and K-Nearest Neighbors. The results highlight that user location, book genre, and publication year slightly affect ratings, with distinct preferences varying by U.S. state and age group. However, the models aren't correlated enough to make good recommendations, hence we recommend focusing on promoting the ten most popular books to enhance user satisfaction and engagement.

2. Introduction

Online bookstores can optimize inventory and increase sales through recommendation systems. This report will cover the key data preprocessing steps performed prior to analysis and how these can be achieved through machine learning.

In recent years, the bookstore industry has undergone a major digital transformation, characterized by a shift from physical stores to digital platforms (Chen et al., 2021). This changed the way books are purchased and the way consumers value and evaluate books. Online ratings have become a key aspect of the digital bookstore experience and are critical for bookstore managers to understand and remain competitive in a market driven by digital consumer behavior.

This report aims to analyze datasets of online bookstores, particularly those located in the United States, to reveal key characteristics that significantly influence user ratings of books. As such, the purpose of this analysis is to guide bookstore managers in making strategic decisions about which books to stock and which to recommend in order to improve sales and customer engagement.

The analysis is based on the following research question: "What key characteristics have the greatest impact on user book ratings in U.S. online bookstores?" Our team went through three main data sources and further preprocessed them by selecting only the features we felt were important in determining user ratings:

- BX-Books: ISBN, author, year of publication and publisher
- BX-Users: User ID, state, country, age
- BX-Ratings: User ID, ISBN, rating

This paper examines the relationship between book features and user ratings. Our findings will guide bookstore managers to develop better strategies for improving and selling efficiency.

3. Methodology

3.1 Tools and Libraries

Our project utilizes various Python libraries, each of which contributes to different aspects of data processing and analysis:

- **Pandas:** provides efficient data frame handling and preprocessing capabilities
- **NumPy:** supports numerical data operations
- **Matplotlib and Seaborn:** used for visualizing data, helping to uncover potential patterns and distributions through graphical representations
- **Natural Language Toolkit (NLTK):** a library for text preprocessing tasks. We use its features to tag, remove stop words, and dry text data, which are basic steps for cleaning text data for analysis
- **Scikit-learn**: this library provides machine learning tools, including utility programs for feature extraction, such as TF-IDF vectorizers for text data. It is also used for encoding categorical data and preparing it for machine learning models.
- **Counter and String**: these two Python modules are used for counting elements and string operations, respectively, to help process categorical data.

3.2 Data Preparation

We perform data preprocessing on three main datasets, BX-Books, BX-Users and BX-Ratings, and use Python for data processing and analysis. Specific presets tailored to each dataset

- **Authors and publishers:** Apply case folding to enforce consistency between these fields by converting all characters to lowercase letters.
- **Year of publication:** The year of publication is normalized by clipping the values between 1950 and 2024
- **City:** The "User-City" column is removed due to excessive granularity, which may lead to overfitting of the prediction model.
- **State:** The shortened state code, if present, is expanded to its full name using a mapping based on standard US state names.
- **Country:** Remove extra characters (e.g., quotes) from the 'User-Country' field. The dataset was then filtered to include only users from the United States.
- **Age:** Clean up the "User-Age" field by replacing non-numeric or missing values with -1 and removing unreasonable ages (e.g., ages greater than 100).

Encoding and merging: Label encoding is used to encode the states of users, authors, and publishers to transform the categorical data into a machine-readable form. After preprocessing, the data sets are combined into a unified frame according to "User-ID" and "ISBN".

3.3 Data Analysis and Interpretation

In this report, we used two main machine learning models, the decision tree and the K-Nearest Neighbor (KNN) classifier, along with a number of analytical techniques to determine the factors that influence user book ratings. The analysis began with a correlation analysis using a heatmap visualization to identify potential predictors of book ratings. This makes our evaluation convenient

Feature importance is evaluated by decision tree analysis, and the importance of each feature is determined by visually showing how different attributes affect the prediction and ranking their importance. KNN analysis is also used to examine the impact of each feature on the accuracy of the model, identifying those features that are important for the predictive power of the model.

Model evaluation and validation involved implementing decision trees and KNN classifiers to predict book ratings, evaluating their performance using cross-validation techniques to ensure the effectiveness and generalization ability of the model. These models are rigorously evaluated using performance

metrics such as precision, F1-score, confusion matrix, and classification reports, which provide a detailed account of the model's precision and recall in different rating categories.

4. Data Exploration and Analysis

Preprocessing part

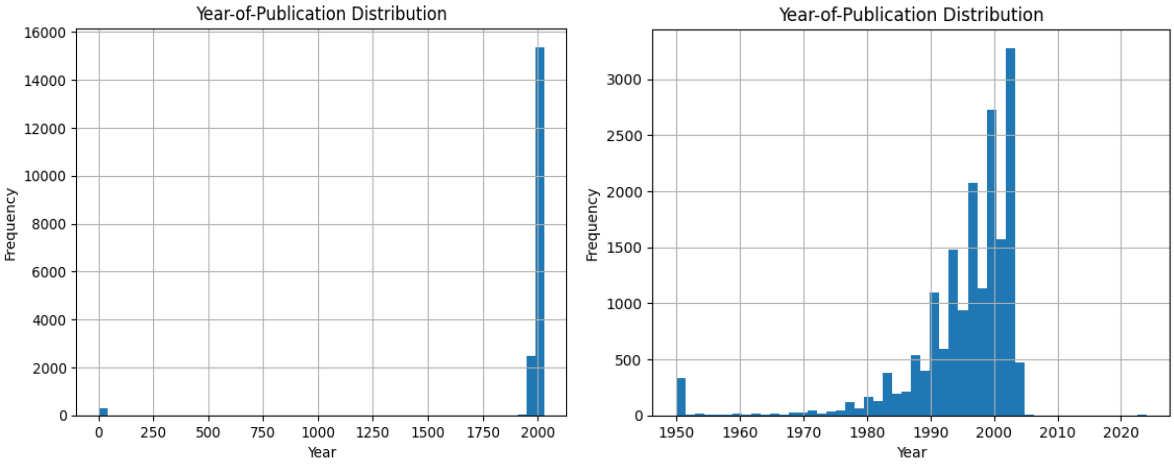

Figure 1

We initially checked the year the book was published and found some outliers, so we narrowed the range down to 1950 to 2024.

After doing some work on the title (such as removing stop words and symbols), we counted the 15 most common words for reference

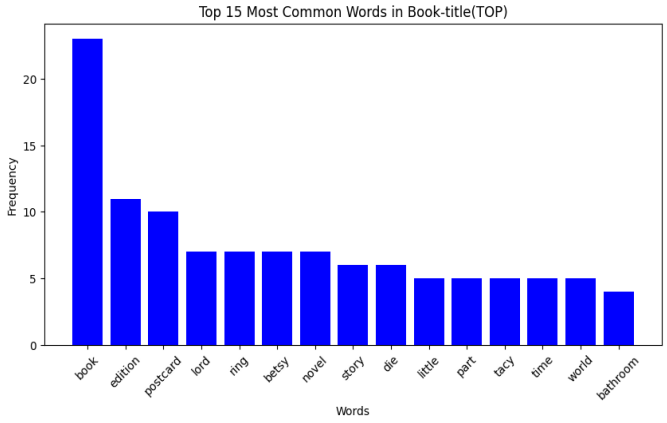

Figure 2

Similarly, we find the most popular authors, publishers, and least popular authors and publishers in the USA

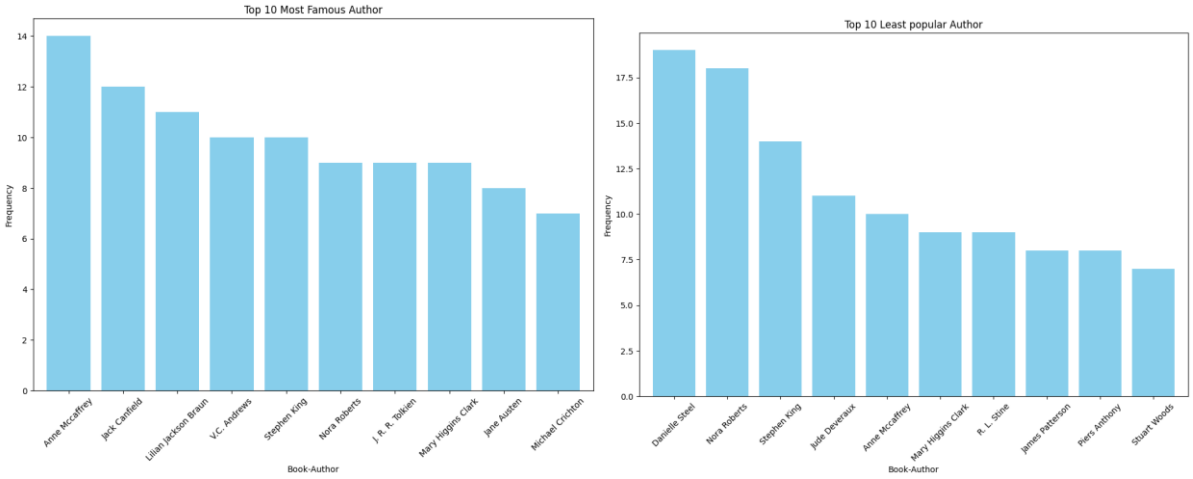

Figure 3

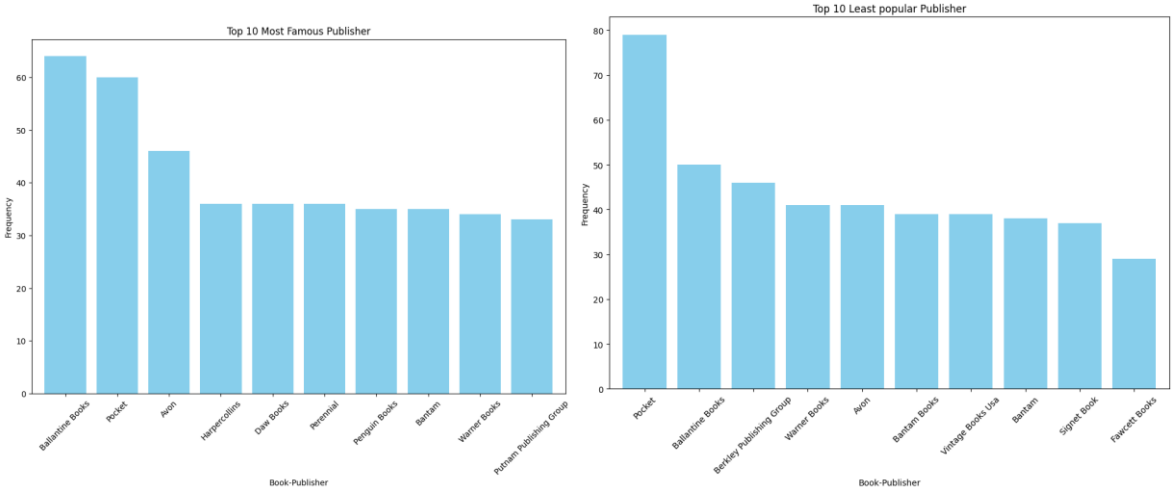

Figure 4

In this data processing script, we first load three datasets: ratings, books, and users. We then cleaned up the user data by removing columns with many missing values, removing extra Spaces and quotes from the string data, and handling invalid state and country information.

Next, we focused on filtering out users located in the United States, and validated and standardized the state names. These processed user data are merged with the user rating data to obtain rating data for USA users

```
# Counting occurrences of invalid state codes
top_invalid_states = invalid_states['User-State'].value_counts().head(5)
print(top_invalid_states)
print()
'''dc          119
ca            8
tx            4
calabria      4
campania      4'''
```

Figure 5

In this part, we found that "dc" had more data, so we changed it to the District of Columbia instead of deleting. This data has since been merged with the book information data by ISBN to form a more comprehensive dataset.

Finally, additional processing was applied to specific fields in the dataset, including range normalization of years, formatting of author and publisher names, and label encoding for subsequent data analysis and

modeling use. The final data were reordered and saved as CSV files, providing a ready dataset for data analysis and reporting.

This process shows how key information can be extracted, cleaned and merged from raw data for detailed data analysis. Such processing not only ensures the consistency and availability of the data, but also lays a solid foundation for possible data analysis projects.

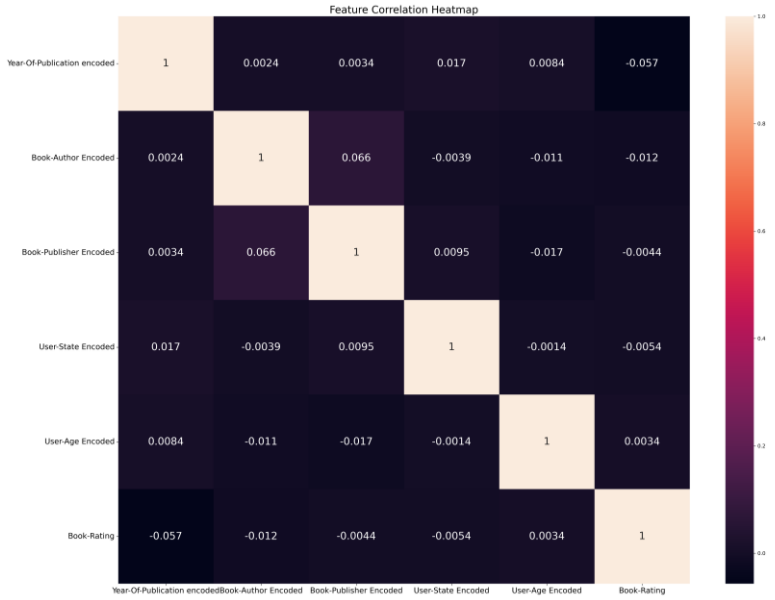

Figure 6

The features correlation heatmap displays a visual representation of the relationships between several encoded variables related to books and user demographics, alongside their correlation coefficients with book ratings. In this heatmap, the diagonal values are uniformly set to 1 which reflects the perfect correlation of each variable with itself, which is standard for any correlation matrix.

Focusing on the correlation values with Book-Rating, it's evident that none of the variables show a strong linear relationship with the rating outcome. For instance, Year-Of-Publication encoded displays a slightly negative correlation of -0.057 with Book-Rating, suggesting that newer books might have lower ratings, although the effect is very subtle. Similarly, other variables like User-State Encoded and User-Age Encoded also show minimal correlation coefficients of -0.0054 and 0.0034 respectively, indicating almost negligible linear relationships with book ratings.

Among the variables themselves, the correlations remain generally weak, with Book-Author Encoded and Book-Publisher Encoded showing a correlation of 0.066, which is one of the higher values observed. This might suggest a slight association between the authors and their publishers, potentially due to specific publishers predominantly working with certain authors. Other notable correlations include Year-Of-Publication encoded with User-State Encoded, which is 0.017, still indicating a very weak relationship. However, the predominantly dark tones across the heatmap underscore the overall lack of strong linear relationships among the features studied and with the book ratings.

This pattern highlighted by the heatmap suggests that predictive modeling for book ratings might require more complex methods capable of capturing deeper patterns not evident through simple correlation, or perhaps the inclusion of more predictive features beyond those currently encoded in the dataset.

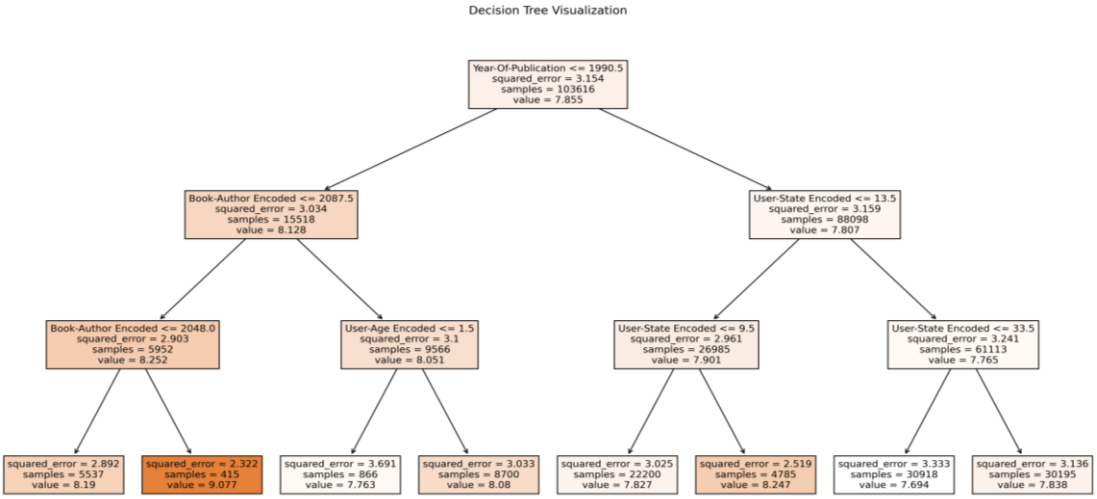

Figure 7

Based on Figure 7, the root node of the decision tree is based on the Year-of-Publication feature, where it tests if Year-of-Publication <= 1990.5. This node includes a large sample size of 103,616 with a square error of 3.159 and average value of 7.855. This split suggests that the decision tree considers the publication year as one of the significant features in determining book ratings to books published before 1990 and those after.

From the root node the tree splits into two primary paths, the left and the right branch. The left branch is the next decision point based on Book-Author Encoded <= 2087.5 which further divide into Book-Author Encoded and User-Age Encoded. The right branch involves split based on User-State Encoded <= 13.5 which continue to divide into User-State Encoded and User-State Encoded.

However, since Book-Author Encoded and User-State Encoded are categorical data which are transformed to numerical data, it may lead to the loss of the literal meaning. Therefore, from figure 7, the only feature that we can take which significantly affects the rating is the year of publication.

```
Classifier: Decision_Tree, Test Size: 0.3, CV: 5
Mean Accuracy: 0.19857142857142857
Mean F1-score: 0.12099845075552138
Classification Report:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00         1
           2       0.00      0.00      0.00         1
           3       0.00      0.00      0.00         5
           4       0.00      0.00      0.00         3
           5       0.14      0.09      0.11        35
           6       0.00      0.00      0.00        24
           7       0.23      0.24      0.23        50
           8       0.30      0.34      0.32        77
           9       0.17      0.15      0.16        54
          10       0.16      0.18      0.17        50

    accuracy                           0.19       300
   macro avg       0.10      0.10      0.10       300
weighted avg       0.19      0.19      0.19       300
```

Figure 8

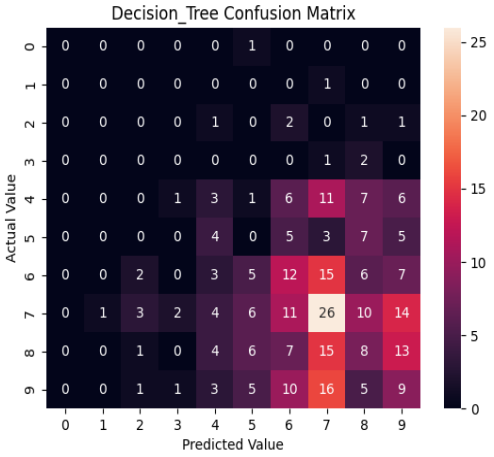

Figure 9

```
Classifier: KNN, Test Size: 0.3, CV: 5
Mean Accuracy: 0.20285714285714285
Mean F1-score: 0.1333279911427513
Classification Report:
              precision    recall  f1-score   support

           1       0.00      0.00      0.00         1
           2       0.00      0.00      0.00         1
           3       0.33      0.20      0.25         5
           4       0.00      0.00      0.00         3
           5       0.15      0.14      0.14        35
           6       0.00      0.00      0.00        24
           7       0.23      0.30      0.26        50
           8       0.27      0.32      0.29        77
           9       0.10      0.07      0.09        54
          10       0.33      0.22      0.27        50

    accuracy                           0.20       300
   macro avg       0.14      0.13      0.13       300
weighted avg       0.20      0.20      0.20       300
```

Figure 10

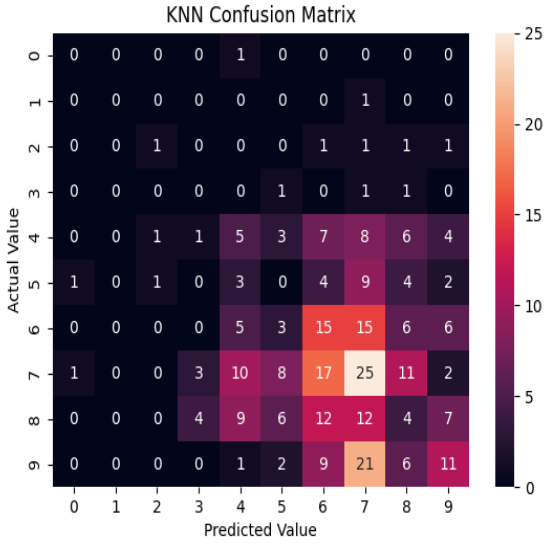

Figure 11

The performance metrics for both the Decision Tree and KNN classifiers show low accuracy, with the Decision Tree at 19.85% (from Figure 8) and the KNN slightly better at 20.86% (from Figure 10). Additionally, both models have notably low F1-scores, with the Decision Tree at 12.09% (from Figure 8) and the KNN at 13.33% (from Figure 10), indicating substantial issues with precision and recall that affect their classification effectiveness.

From the classification reports in Figure 8 and 10, both models exhibit challenges with precision and recall across almost all rating classes, with slight improvements in higher rating classes. This suggests a relative strength in identifying more distinct or common ratings, despite overall poor performance. Furthermore, variability in class support indicates that biases towards more frequently occurring ratings may be influencing the learning outcomes for both classifiers.

The confusion matrices in Figure 9 and 11 analyses for both classifiers reveal low values in the main diagonal, confirming their struggles with accurate predictions. There are significant misclassifications, particularly between closely spaced rating categories, highlighted by a notable spread of values in off-

diagonal cells, especially between consecutive ratings. This pattern suggests both models approximate the vicinity of a rating rather than accurately pinpointing it.

General observations and strategic considerations point to poor generalization capabilities of both models, with potential overfitting in the Decision Tree and underfitting or improper parameterization in the KNN. The features used seem inadequate for capturing essential nuances for accurate predictions, necessitating refined feature engineering or selection. Additionally, uneven class distribution suggests data imbalance issues, which might be addressed with resampling or synthetic data generation. Model-specific enhancements such as simplifying the Decision Tree or tuning parameters for the KNN could potentially improve their predictive performances.

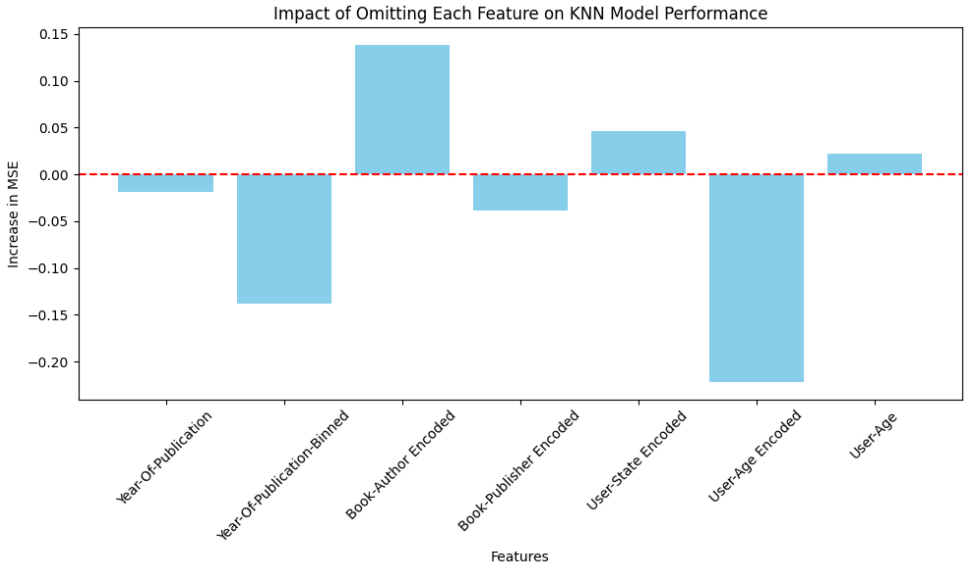

Figure 12

The graph illustrating the impact of omitting specific features on the Mean Square Error (MSE) in a KNN model used to predict user ratings provides insightful revelations about the significance of various features. The binned publication year, while showing a decrease in MSE upon omission, suggests that while the publication year carries some weight, its impact is less critical compared to other features. In stark contrast, the author of a book emerges as a pivotal predictor, with the omission leading to a noticeable increase in MSE, reflecting user biases or preferences toward certain authors. Similarly, the publisher of a book significantly influences model predictions, emphasizing the role of publisher reputation and user preferences. Additionally, the encoded user state substantially affects the model's performance, likely reflecting regional book availability or cultural preferences. Lastly, user age proves to be an important demographic factor; its exclusion results in a slight increase in MSE, indicating its relevance in aligning with user preferences for certain types of books, topics, or complexity levels. Collectively, these insights underscore the nuanced roles that different features play in shaping the predictive accuracy of the model, highlighting areas that could benefit from focused data collection and feature engineering to enhance model robustness and prediction accuracy.

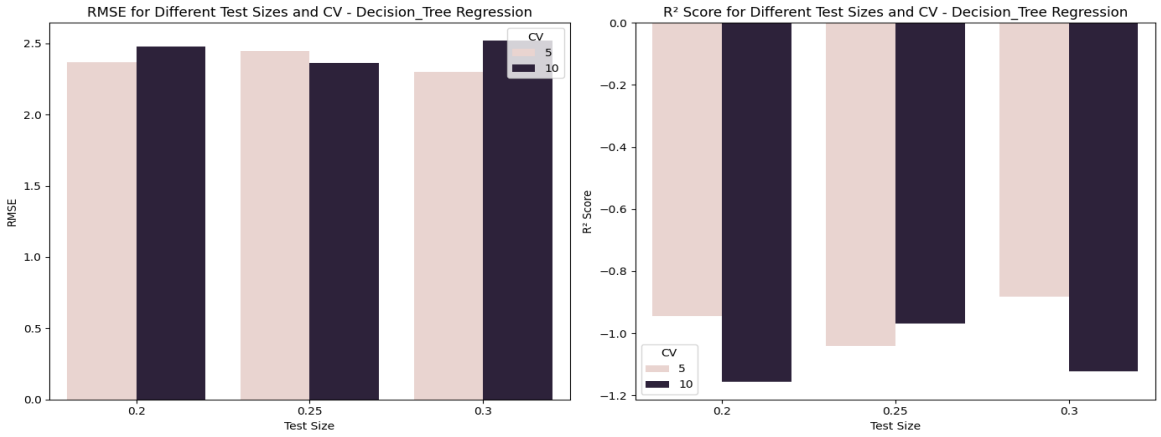

Figure 13

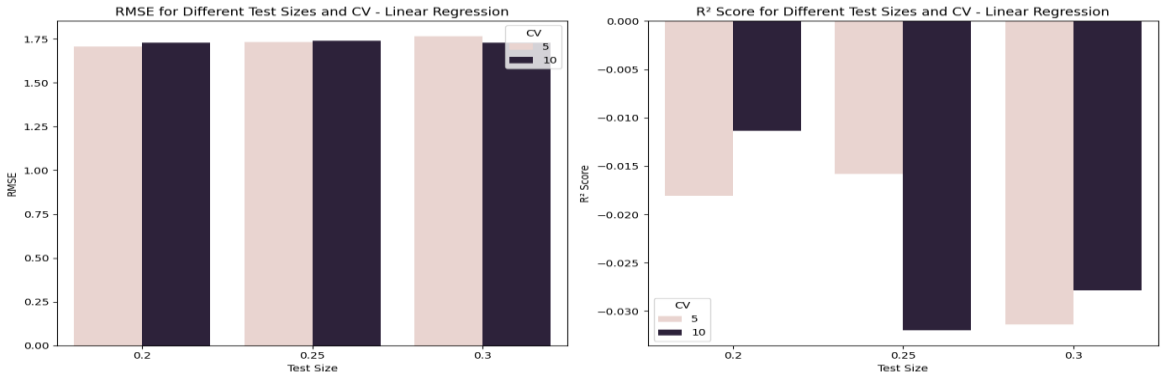

Figure 14

The analysis delves into the cross-validation results of linear regression and decision tree regression models on a dataset. In linear regression, as the test size increases from 0.2 to 0.3, RMSE decreases, suggesting improved stability with a more diverse dataset. However, all configurations yield negative R^2 scores, indicating poor model fit. This implies either model inadequacy or insufficient features capturing the data nuances.

For decision tree regression, similar to linear regression, RMSE decreases with larger test sizes, albeit with significantly higher errors compared to linear regression. Moreover, the 10-fold cross-validation produces generally higher errors, hinting at less consistency compared to the 5-fold cross-validation. Negative R^2 scores persist and worsen with larger test sizes, further emphasizing the model's unsuitability for the dataset.

Consequently, both models exhibit poor fit and predictive performance, necessitating reevaluation of features and modeling approaches. Recommendations include incorporating additional features, exploring advanced regression techniques like ridge or lasso, and improving data quality. Such strategies aim to enhance predictive accuracy and ensure more robust modeling strategies in this context.

5. Results

There is a slight correlation (0.91) between User-Age and User-Age Encoded, which is expected because: Results: Insights from Supervised Learning Models

This research is more focus on conducting an experiment with supervised learning models by analyzing models with classifiers such as decision trees and KNN, to linear and decision tree regressions, that provided useful insight into the features that influence user ratings of books on an online bookstore. The

main emphasis of the results is on the efficacies of the models in predicting ratings based on different features and their relations with user preferences.

Both the Decision Tree and KNN classifiers had an accuracy close to 20%. Such a low accuracy points to a big challenge in determining an accurate prediction of user ratings exactly, and it implies either an issue in model adequacy or feature selection. The confusion matrices on these classifiers revealed quite some confusion between adjacent rating categories, meaning it had a hard time distinguishing between similar books. For instance, books with rating 7 were pretty often misclassified in either class 6 or 8. From the analysis, it is also shown that features like 'Book-Author Encoded' and 'User-State Encoded' highly affect the performance of the models when present or not. This is a clear indication of the importance of authorship or user regional preferences with respect to how books are rated.

The performance of regression models, including both Linear and Decision Tree models, was negative with test sizes and different cross-validation configurations, resulting in very low and quite poor R^2. It suggests that the models do not generalize the variance in user ratings. RMSE metrics further unveiled the inaccuracy of the models, suggesting high error rates in prediction for all test sizes and cross-validation configurations. Those metrics detail how high levels of predictive accuracy are hard to be achieved, which in turn implies the need for more sophisticated models or more sophisticated feature engineering.

Exemplary here are the Feature Omission Impact Graphs that reveal a drastic rise in MSE upon the omission of the key features, among which the 'Book-Author Encoded' and 'User-State Encoded' turned up to play a crucial role within the models of prediction. Furthermore, the attached confusion matrices provide a visual idea of the domains where the models were successful and where they failed in general, showing specific challenges in the aspect of rating prediction.

In other words, this analysis has made a strong effort toward better understanding of the dynamics of predictiveness of user book ratings in an online bookstore. Low performances of both classifiers and regressors suggest that the task is tough and user ratings are most probably dependent on interaction among known features, e.g. authors and publishers, and unknown or latent factors that the current models were not able to capture. This is a very important finding for further model refinements and adjustments in feature sets that would make it more suitable with respect to user preferences, resulting in increasing accuracy in making predictions. Such an analysis aids not only in understanding the effectiveness of current predictive models but also in directing future efforts of data collection, feature engineering, and model selection that can be channeled toward making improvements in the recommendation systems of the bookstore and, hence, toward enhancing the satisfaction of its users.

6. Discussion and Interpretation

Effect of Author and Publisher on Ratings The results showed that 'Book-Author Encoded' and 'Book-Publisher Encoded' have significant effects on the model predictions. Inclusion of these features drastically increases MSE when omitted from the features. This shows that people develop preference or bias towards some authors and publishers, which could be by their quality, genre, or simply satisfaction with their works. Realization of such fact, therefore, shows that in the publishing world, reputation and brand loyalty of a particular author and publisher influence user choice and satisfaction as well. In case the authors and publishers have such an impact, then it means that a book store can structure his or her stock and marketing based on those respected authors, and this will boost user interest and sales.

Regional and Demographic Influences on User Ratings An interaction term between the features 'User-State Encoded' and 'User-Age' has an important effect on the prediction errors, thereby indicating that geographical and demographic factors impact the way in which users rate books. This latter can be read in terms of discrepancies in respect to book availability, cultural tastes in books, or age-related genre or theme preferences. Such complexity, as pointed out by the variance in demographic and regional

characteristics, hints toward a very nuanced relationship between a user's background and his behavior of rating a book. These are invaluable insights toward the customization of user experience and recommendations for book offerings, thereby increasing user satisfaction and engagement based on regional trends and demographic preferences.

The classifiers and regression models showed low metrics on almost every aspect: precision and recall, with markedly negative R^2 scores. This kind of behavior could be indicative that the features now under use cannot encapsulate the complete gamut of considerations that users employ in rating books, or, in other words, the possible non-linear relationships or complex interactions between features that linear models are not able to capture. Performance problems have to be underlined so that future modeling efforts are oriented to complex models, and the feature set enriched with the ability to reflect those dynamics more accurately.

Very Stable Feature Impact Across Model Configurations The impact of important features, such as 'User-Age' and 'Book-Author Encoded', is very consistent across model configurations and different test sizes. Such stability would argue that such features are key to the representation of user rating behavior, hence substantially robust to changes in model parameters. The influence of these features being stable across different settings gives a solid base for incrementally developing predictive models, whereby future improvements should focus on facets of data that continuously influence outcomes.

Negative R^2 Scores in Regression Models All regression models returned unexpectedly negative values of R^2, meaning that they were worse than the model predicting the mean rating for all observations. Probably, it is indicating a problem of overfitting or the unsuitability of the regression models specified for complexity and data structure. This is very important for the recognition that the regression approaches taken herein may not be appropriate to the problem or data structure under inquiry. It therefore stimulates the development of alternative models or methodologies, such as ensemble methods or advanced machine learning techniques, able to handle high-dimensional spaces and interaction effects more effectively, possibly increasing predictive accuracy and reliability.

7. Limitations and Improvement Opportunities

Dataset Limitations

- Data coverage: Our dataset is mostly limited to users in a specific geographical location (USA) and a certain age group. This may limit the model generalization ability as it may not be fully applicable to users in other regions or age groups.
- Sample size and bias: Certain classifications, such as certain states or age groups, may not be adequately represented in model training due to insufficient number of samples.

Model Limitations

- Risk of Overfitting: Especially in decision tree models, there is a risk of overfitting, making the model too dependent on specific sample features in the training data and not able to generalize to new data.
- Feature selection and processing: Current feature processing may not be ideal. For example, user age and user status are directly encoded without more nuanced grouping or transformation, which may affect the interpretability and accuracy of the model.

Opportunities for Improvement

- Data Collection and Expansion: The dataset is expanded to improve the generalization and accuracy of the model.
- Feature Engineering: Further research and implementation of more sophisticated feature engineering techniques can help reveal deeper patterns in the data.
- Model Optimization and Integration: Explore more advanced algorithms and model integration methods to improve prediction accuracy and stability.

8. Conclusion

Taken together, user-state Encoded, user-age, and book-author Encoded are the key features that affect User Book ratings in online bookstores in the United States. The user-age is the most important feature. These findings are particularly important for bookstore managers not only to help understand consumer preferences, but also to guide inventory choices and marketing strategies to appeal to specific groups of readers. For example, to promote a specific genre of books based on state-by-state data, or to develop age-specific promotions based on age data. However, we finally found that this prediction result has some correlation, but it is not significant. We decided just to give the top 10 most famous books. Through an in-depth analysis of these key characteristics, bookstores can remain competitive and relevant in a highly competitive market.

9. References

[1] Chen, C. M., Bao, S. L., Feng, T., Lu, Y. T., & Li, R. (2021, December). Under the Prevalence of E-Commerce: Online Bookstore System. In *2021 9th International Conference on Orange Technology (ICOT)* (pp. 1-5). IEEE.